Recueil planifié des données: plans d'expérience et d'échantillonnage Introduction

Philippe Letourmy
Cirad, novembre 2017
UR Agroécologie et intensification durable des Cultures Annuelles

2 étapes sont cruciales dans toute étude statistique:

1. Recueil des données
2. Analyse de ces données

Souvent l'accent est mis sur l'ensemble des méthodes d'analyse statistique des données, mais la question des méthodes de recueil est souvent négligée. Elle est souvent associée aux seules méthodes pour lesquelles elles ont fait l'objet de développements conjoints: l'analyse de variance pour les dispositifs expérimentaux et les estimations de moyennes ou de proportions pour l'échantillonnage. Mais toute étude statistique passe par un recueil de données.

Expérience et échantillonnage

Expérience :

- Possibilité de maîtriser les facteurs à étudier
- Comparer des traitements, « toutes choses étant égales par ailleurs »
- Résultats relatifs à une situation particulière (conditions d'expérience)

Échantillonnage :

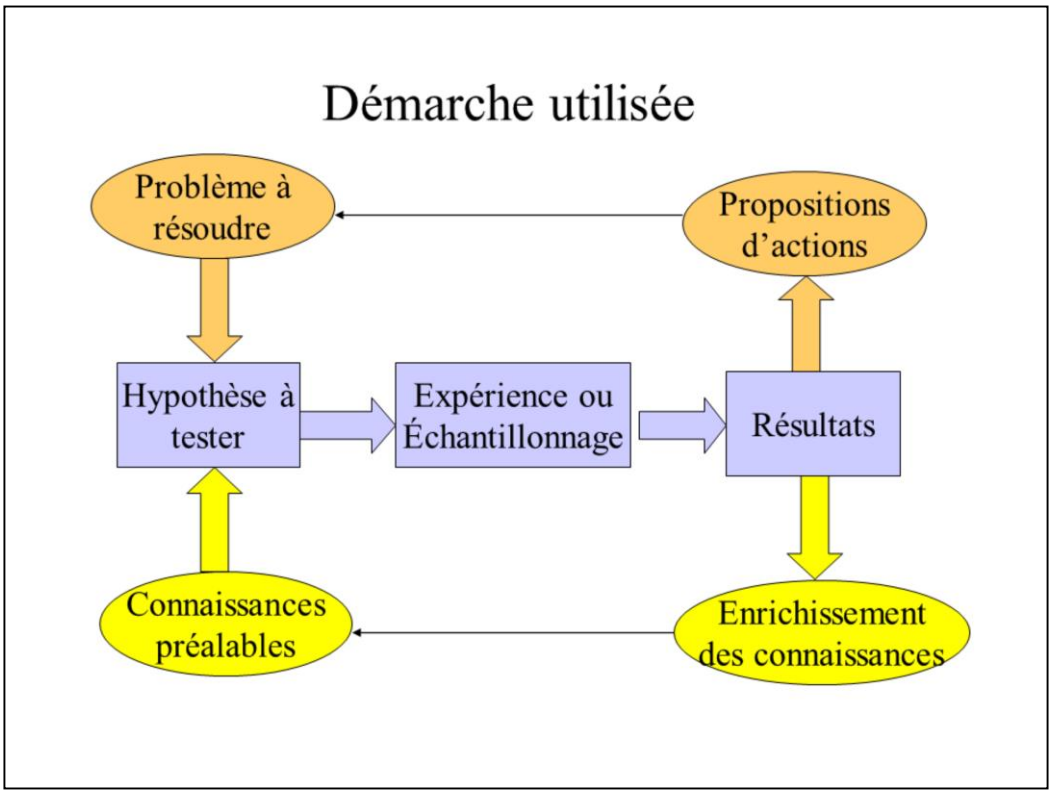
- Observer des unités d'une population
- Représenter la population par une partie de celle-ci (l'échantillon)
- Résultats utilisables si la représentativité est assurée

Tout recueil d'informations est conçu en fonction d'un objectif, et cet objectif conditionne la manière dont seront collectées les données et les questions auxquelles on tentera d'apporter une réponse.

On considère deux modes de recueil planifié de l'information : l'expérimentation et l'échantillonnage.

L'expérience recherche analytiquement des relations de cause à effet.

L'échantillonnage recherche la généralisation de conclusions d'observations à toute une population.



Les problèmes rencontrés (en agriculture, en médecine, dans l'industrie…) et les connaissances préalables concourent à la définition d'une hypothèse à tester par la planification du recueil de données, ce qui aboutit à des résultats, sous la forme d'une hypothèse nouvelle. Ces résultats débouchent d'une part sur des propositions d'actions concrètes, d'autre part sur un enrichissement des connaissances, qui peuvent aussi être utilisées pour résoudre des problèmes similaires.

Vocabulaire de l'expérimentation (1)

- Facteur : toute série d'éléments de même nature conditionnant le phénomène étudié (facteur étudié/de contrôle, qualitatif/quantitatif)
- Niveau ou modalité : un des éléments qui constituent un facteur
- Objet ou traitement : toute combinaison de niveaux ou de modalités de tous les facteurs étudiés
- Unité expérimentale : unité élémentaire qui reçoit un traitement et sur laquelle est faite chaque mesure

Les facteurs, qui sont l'objet même de l'expérience, sont appelés facteurs étudiés. Ceux qui sont liés à la variabilité du milieu et introduits de façon à ce que leurs effets puissent être éliminés sont appelés les facteurs de contrôle (blocs).

Exemples de facteurs étudiés: la variété, la dose d'engrais, le médicament, etc.

Il faut noter que, pour éliminer des effets de bordure, on peut être amené à distinguer la parcelle utile sur laquelle sont effectuées les mesures et la parcelle expérimentale (= parcelle utile + bordures) qui reçoit le traitement. Cela peut se généraliser à toute unité, dès lors que l'on distingue l'unité sur laquelle est faite la mesure et celle qui reçoit le traitement.

Il faut aussi remarquer que la mesure sur chaque unité expérimentale peut être faite à partir d'un échantillonnage.

Vocabulaire de l'expérimentation (2)

- Champ d'expérience : ensemble des unités expérimentales
- Champ d'application : domaine sur lequel les résultats sont applicables
- Observation : mesure de caractère quantitatif ou qualitatif effectuée sur chaque unité expérimentale
- Plan d'expérience : il définit la manière dont les traitements sont affectés aux unités expérimentales

Chaque observation est entachée d'une erreur expérimentale, ou erreur résiduelle, somme de l'erreur unitaire et de l'erreur technique. On appelle erreur technique, ou erreur de mesure, l'erreur que l'on fait lors de l'observation. Cette erreur doit être aussi petite que possible pour avoir des mesures fiables et précises, et par là avoir une expérimentation avec des résultats interprétables. On appelle erreur unitaire l'erreur due à l'hétérogénéité des unités expérimentales (quand bien même on y appliquerait le même traitement), appelée aussi erreur d'hétérogénéité. On l'appelle aussi erreur de randomisation, car c'est la randomisation qui en fait une variable aléatoire dont on peut étudier la distribution.

Vocabulaire de l'échantillonnage (1)

- Population : ensemble des individus étudiés (N = taille de la population)
- Échantillon : sous-ensemble observé de la population (n = taille de l'échantillon)
- Unité statistique : unité sur laquelle est faite chaque mesure
- Base de sondage : liste exhaustive des unités de la population

Le choix de la population étudiée est important, car c'est tout l'objet de l'échantillonnage de connaître cette population et il conditionne le domaine sur lequel les résultats sont applicables.

L'échantillon résulte d'un tirage, aléatoire ou non aléatoire.

f = n / N est le taux de sondage

Vocabulaire de l'échantillonnage (2)

- Recensement : échantillon particulier comprenant toute la population
- Panel : échantillon sur lequel on effectue des mesures répétées dans le temps
- Représentativité : tout individu peut figurer dans l'échantillon avec une probabilité non nulle et connue
- Plan d'échantillonnage : il définit la manière dont les unités de la population sont choisies

Comme dans les expérimentations, l'observation est entachée d'erreur; celle-ci est la somme de l'erreur de mesure et de l'erreur d'échantillonnage. L'erreur de mesure est l'erreur que l'on fait lors de l'observation et l'erreur d'échantillonnage est une erreur due au choix de l'échantillon particulier et à l'hétérogénéité des individus de la population étudiée.

Le protocole de recueil des données répond à des objectifs préalablement fixés

- Le point de départ est un modèle, si possible formalisé, du phénomène étudié
- Ce modèle amène à formuler clairement les questions auxquelles le recueil des données doit répondre (partie la plus importante)
- Le plan d'expérience ou d'échantillonnage doit être cohérent avec les questions posées, voire même optimisé pour obtenir la meilleure précision

1) Introduction. Expérimentation planifiée

- Les 3 principes de l'expérimentation
  - Randomisation : l'allocation des traitements aux unités est faite par un tirage aléatoire
  - Répétition : chaque traitement est affecté à plusieurs unités, afin de pouvoir estimer une erreur expérimentale
  - Contrôle de l'erreur : pour réduire la part non contrôlée de l'expérience, donc diminuer l'erreur expérimentale

Pourquoi la randomisation ? tout d'abord elle permet d'éviter tout biais plus ou moins conscient. En termes mathématiques, elle permet d'assurer que l'erreur unitaire est d'espérance nulle. Il est important de savoir que la randomisation n'annule pas, ni même ne diminue l'erreur. Elle permet d'éviter qu'elle soit systématique et de connaître suffisamment sa distribution pour assurer la validité du test des effets traitement.

Pourquoi des répétitions ? Pour permettre d'évaluer la part de l'erreur expérimentale dans l'observation. En effet, sans des répétitions de chaque traitement sur plusieurs unités, on ne peut pas distinguer l'effet dû au traitement de l'erreur expérimentale.

Enfin, il est nécessaire d'avoir une erreur expérimentale aussi petite que possible, c'est-à-dire de contrôler l'erreur : choisir un champ expérimental aussi homogène que possible, former des groupes homogènes de parcelles, ou blocs, lorsqu'une hétérogénéité est reconnue sur le terrain, et essayer de minimiser l'erreur de mesure.

- La randomisation

Soit U l'ensemble des unités, numérotées de 1 à N
Soit T l'ensemble des traitements numérotés de 1 à t

Un plan expérimental est alors la donnée d'une application S: U→T

Un plan randomisé est la donnée d'une loi de probabilité uniforme sur un sous-ensemble des applications S

U est aussi appelé le champ d'expérience.

A chaque type de plan d'expérience correspond un sous-ensemble de toutes les applications S

- Plan en randomisation totale à n répétitions

On considère que N=t*n (plan équilibré)

Soit une bijection de U→T x {1…n}

Un plan en randomisation totale est alors donné par un tirage aléatoire parmi l'ensemble des permutations dans U={1…N}

Exemple avec 5 traitements et 3 répétitions:

T2	T3	T2	T1	T2
T3	T4	T1	T1	T5
T4	T3	T5	T5	T4

• Modèle de randomisation (1)

Soit m_{iu} la réponse conceptuelle de l'unité u au traitement i
On s'intéresse à $t_i = m_{i*}$ (moyenne du traitement i)
Une observation de la répétition k du traitement i s'écrit:

$Y(i,k) = \Sigma_u \delta_u(i,k) m_{iu} + \varepsilon(i,k)$
Où $\delta_u(i,k) = 1$ si $s(u)=(i,k)$ et 0 sinon
$\varepsilon(i,k)$ est l'erreur de mesure sur $Y(i,k)$

On pose $m_{iu} = t_i + (m_{iu} - t_i) = t_i + d_u$ (additivité des effets traitement et unité)

$\delta_u(i,k)$ est une variable aléatoire qui prend les valeurs 1 ou 0 selon que la permutation aléatoire affecte ou non à l'unité u

la kième répétition du traitement i

• Modèle de randomisation (2)

Alors:

$Y(i,k) = t_i + \Sigma_u \delta_u(i,k)\, d_u + \varepsilon(i,k)$

Ou encore:

$Y(i,k) = t_i + e(i,k) + \varepsilon(i,k)$

e(i,k) est l'erreur unitaire, ε(i,k) l'erreur de mesure, l'erreur expérimentale est donc la somme des 2

On a:

$E(Y(i,k)) = t_i$ $\quad$ $Var(Y(i,k)) = (N-1)^* \sigma_u^2 /N + \sigma_e^2$

$Cov(Y(i,k), Y(j,l)) = - \sigma_u^2 /N$ si $(i,k) \neq (j,l)$

On appelle aussi e(i,k) l'erreur d'hétérogénéité et ε(i,k) l'erreur technique; la somme des 2 est aussi appelée l'erreur résiduelle.

• Modèle de randomisation (3)

On a pour comparer 2 moyennes:

$\mathrm{Var}(Y(i,.) - Y(j,.)) = 2\,(\sigma_u^2 + \sigma_e^2)/n$

Et asymptotiquement (quand n devient grand):

$Y(i,k) = t_i + E(i,k)$

converge vers un modèle gaussien d'analyse de variance à 1 facteur, avec $E(E(i,k))=0$ et $\mathrm{Var}(E(i,k))= \sigma_u^2 + \sigma_e^2$ et les erreurs indépendantes entre elles

Le modèle d'analyse de variance est asymptotiquement atteint, mais les contrastes ont des variances estimées de façon non biaisée, même si n reste petit.

• Puissance et nombre de répétitions (1)

Combien de répétitions faut-il réaliser pour mettre en évidence une différence Δ entre les deux traitements avec une probabilité $1-\beta$ (puissance du test) dans un test au niveau α ? Mais, pour y répondre, il faut encore une autre information : la valeur de la variance résiduelle de l'essai.

Si $\mu + \alpha_1$ est la moyenne du traitement 1 et $\mu + \alpha_2$ la moyenne du traitement 2, on teste $\alpha_1 = \alpha_2$ par la statistique de Student

$$t(dl) = \frac{A_1 - A_2}{s\sqrt{2/n}}$$

Soit le modèle $Y_{ij} = \mu + \alpha_i + E_{ij}$ on nomme A_1 et A_2 les estimations respectives de α_1 et α_2

s est une estimation de l'écart-type résiduel σ

le risque α est une valeur préalablement fixée, alors que le risque β est une fonction des différences entre traitements dans le cadre d'une expérience. En fait β (et donc aussi la puissance $1-\beta$) est une fonction :

- du risque α,
- des différences vraies entre traitements,
- de la variabilité résiduelle,
- du nombre de répétitions.

• Puissance et nombre de répétitions (2)

On pose $\alpha_1-\alpha_2=\Delta=E(A_1-A_2)$, donc la statistique suivante suit une loi de Student à dl degrés de liberté

$$\frac{A_1 - A_2 - \Delta}{s\sqrt{\frac{2}{n}}}$$

Or le test rejette l'égalité des traitements si

$$\left|\frac{A_1 - A_2}{s\sqrt{\frac{2}{n}}}\right| > t_{\alpha/2}(dl)$$

On teste l'hypothèse H_0: $\alpha_1 = \alpha_2$ contre l'alternative H_1: $\alpha_1 \neq \alpha_2$

• Puissance et nombre de répétitions (3)

pour mettre en évidence une différence Δ avec une probabilité 1-β dans un test au niveau α, il faut :

$$\frac{\Delta}{s\sqrt{\frac{2}{n}}} \geq t_{\alpha/2}(dl) + t_{\beta}(dl)$$

donc le nombre de répétitions minimal pour ceci est :

$$n = 2\frac{s^2}{\Delta^2}\left(t_{\alpha/2}(dl) + t_{\beta}(dl)\right)^2$$

En fait, on ne peut connaître dl que si on a n. Dans la pratique, on procède de manière récursive, jusqu'à obtenir le nombre de répétitions adéquat respectant la formule ci-dessus.

On peut aussi se servir d'un abaque pour s'éviter des calculs. Un abaque de la fonction puissance du test de Student a été établi pour les niveaux 1% et 5%. La puissance est donnée en fonction de φ (= (Δ√n)/ 2s)

On peut donc déduire la puissance 1-β à partir de n, Δ et s, ou bien le nombre de répétitions n à partir de s, Δ et 1-β.

Exercice d'application : essais de sélection de variétés de canne à sucre.

2) Introduction. Echantillonnage

Démarche en échantillonnage

- Représenter la population par une partie de celle-ci (l'échantillon)
- Formalisme et remarque historique
- Description des types de problèmes à résoudre

Tout recueil d'informations est conçu en fonction d'un objectif, et cet objectif conditionne la manière dont sont collectées les données et les questions auxquelles on tentera d'apporter une réponse.

L'échantillonnage recherche la généralisation des conclusions à toute une population.

Démarche en échantillonnage (1)

- Soit U une population : $\{u_1 \ldots u_N\}$
- Soit le caractère y inconnu, fonction : $u \rightarrow y_u$
- Calculer une fonction du caractère y, $h : y \rightarrow h(y)$ (h connue)
- Échantillon : sous-ensemble observé de la population, $s = \{u_1 \ldots u_n\}$
- On utilise une statistique $T(s,y_s)$ pour estimer h(y)
- On cherche un compromis entre efficacité et coût d'échantillonnage

Pour simplifier, on considère y unidimensionnel, mais réel, entier ou binaire (0 ou 1).

La fonction h est réelle; les exemples classiques sont h(y) = moyenne de y (ce qui correspond à la proportion de 1 lorsque y est binaire), ou bien h(y) = variance de y, mais on peut aussi s'intéresser à la médiane ou à un autre quantile

L'échantillon résulte d'un tirage, aléatoire ou non aléatoire. Si U est fini et que s=U, il s'agit d'un recensement.

L'efficacité mesure la petitesse de l'erreur $T(s,y_s) - h(y)$ et le coût est une fonction croissante de la taille de l'échantillon.

Démarche en échantillonnage (2)

2 sources d'aléatoire sont possibles

- y peut être un champ aléatoire $Y=(Y_u)\ u\in U$, avec $Y_u : \omega_1 \rightarrow Y_u(\omega_1)$, on cherche à estimer $h(Y(\omega_1))$; notons que $Y_u(\omega_1)-EY_u$ (erreur de modèle) peut ne représenter que l'erreur de mesure de y_u
- L'échantillon peut être tiré au hasard, selon un plan d'échantillonnage de loi de probabilité P, alors $s = S(\omega_2)$
- Les données (s, y_s) sont fonction de $\omega=(\omega_1, \omega_2)$

Comme dans les expérimentations, l'observation est entachée d'erreur; l'erreur totale est la somme de 2 erreurs: l'erreur de modèle et l'erreur d'échantillonnage.

$\omega=(\omega_1, \omega_2)$, ω_1 est donné par la nature et ω_2 par le chercheur qui conçoit le plan; les données sont fonction de ω:

$\omega \rightarrow (s=S(\omega_2), Y_s(\omega_1))$

Démarche en échantillonnage (3)

- Si la loi P est la même pour toute réalisation ω_1, alors P est indépendante de M (loi de Y); on dit que le plan de sondage est non instructif pour Y
- Estimer $h(Y(\omega_1)) = Y_{u0}(\omega_1)$, avec $u_0 \in U$ et $\notin s$, est un problème classique d'extrapolation
- La loi M est inconnue, mais une connaissance partielle peut être apportée dans une optique bayésienne ou bien inférentielle classique

Les plans non instructifs excluent les méthodes séquentielles par exemple.

Estimer $h(Y(\omega_1))$ n'est pas un problème d'inférence classique mais un problème de filtrage

La statistique bayésienne pose une loi a priori sur l'ensemble des lois M (cet ensemble connu de lois est appelé le modèle), loi a priori que ne pose pas la statistique classique (même si celle-ci prend en compte le modèle statistique).

Démarche en échantillonnage (4)

Remarque historique

- Au début on a tiré des échantillons non aléatoires « selon le but »; rôle prépondérant du modèle M
- Puis on a insisté sur le choix aléatoire de l'échantillon; rôle prépondérant du plan de sondage P
- Enfin une prise en compte conjointe de P et de M est possible dans le cadre de la théorie de la décision

« selon le but » ou « purposive sampling »: on tient compte de la connaissance disponible sur le caractère par le modèle.

En focalisant sur le plan de sondage, on travaille indépendamment de la distribution de probabilité du caractère observé.

Le choix d'une stratégie (combinaison d'un plan de sondage et d'un estimateur) nécessite de pouvoir évaluer la qualité de la stratégie en mesurant un risque.

Démarche en échantillonnage (5)

- Pour choisir une stratégie (P,T), on part d'une fonction de perte L(y, s, t) du caractère y, de l'échantillon s et de la valeur t de la statistique
- On en tire un risque R(M, P, T) pour un plan non instructif = $E_M E_P (L(Y, S, T))$
- Exemples de fonctions de perte:
  - coût du tirage $L_c(s) = c \#s$, d'où un risque $R_c(P)= c\, n_P$
  - Erreur quadratique d'estimation $L_e(y,t)=(t-h(y))^2$, d'où $R_e(M,P,T)$= erreur quadratique moyenne

Ce qui nous amène aux types de problèmes qui peuvent être résolus dans ce cadre (théorie de la décision):

- 1) calculer les risques $R_c(P)$ et $R_e(M,P,T)$ pour loi M et stratégie (P,T) données,
- 1') comparer les risques de 2 stratégies pour une loi M,
- 2) pour un modèle $\subset\mathcal{M}$ et un plan P donnés, trouver un « bon » estimateur,
- 3) pour $\subset\mathcal{M}$ et un budget B donnés, trouver une « bonne » stratégie (P,T) tq $R_c(P) \leq B$,
- 4) soit une fonction η: $M \rightarrow \eta(M)$ $M \in \subset\mathcal{M}$, on cherche la stratégie (P,T) qui minimise $R_c(P)$ tq $R_e(M,P,T) \leq \eta(M)$.

Démarche en échantillonnage (6)

Pour terminer sur cette introduction à l'échantillonnage et aux sondages, il faut parler de la notion de superpopulation (Ardilly 2006) :

> *Population fictive de taille infinie utilisée dans « l'approche modèle » de la théorie des sondages, et constituant un cadre formel assez pratique pour justifier la manipulation de variables aléatoires de nature « classique » (aléa qui se superpose à l'aléa généré par l'échantillonnage). Dans cette optique, on peut considérer en particulier que la population finie réelle dont on dispose et dans laquelle on échantillonne est elle-même un échantillon aléatoire issu d'une ou plusieurs superpopulations.*

Dans toute la suite, on considèrera une partie des questions soulevées ci-dessus au cas où y n'est pas aléatoire et où U est fini. On se limitera à des plans de sondage où il faut estimer une moyenne ou une proportion. Mais il est important d'avoir à l'esprit que le problème est plus général en réalité.

Nous nous focaliserons plus spécifiquement sur les plans de sondage. La notion de superpopulation se retrouve lorsque nous aborderons les compléments sur l'échantillonnage et les plans basés sur un modèle des données.